Zoned Namespaces in Practice

Matias Bjørling
Director, Emerging System Architectures

August 6th, 2019
What are Zoned Block Devices?

The new paradigm in storage

• The storage device logical block addresses are divided into ranges of zones.
• Writes within a zone must be sequential.
• The zone must be erased before it can be rewritten.
Zoned Storage has already been in HDDs

• SMR (Shingled Magnetic Recording)
  – Enables areal density growth
  – Causes magnetic media to share flash access model
    • Data must be erased to be re-written

• Zoned Access
  – Zoned Block I/F standardized in INCITS
    • Zoned Block Commands (ZBC): SAS
    • Zoned ATA Commands (ZAC): SATA
  – Host/Device cooperate to optimize RMW aspect of SMR by enforcing sequential writes and enabling host FTL model
Why Zoned Storage
Addressing the needs of large-scale data infrastructure

Why Zones for Solid State Drives?
Ubiquitous Workloads

The cloud applies multiple workloads to a single SSD

SSDs write log-structured to the media that requires garbage collection

Multiplex data streams onto the same garbage collection units

Increases Write Amplification, Over-Provisioning and thereby Cost

Decreases throughput and latency predictability
Zones for Solid State Drives

Eliminate data streams multiplexing:
- Significantly decreases write amplification, over-provisioning and thereby reduces cost
- Increases throughput and latency predictability
Zoned Namespaces

• Ongoing Technical Proposal in the NVMe™ working group
• New Zoned Command Set – Inherits the NVM Command Set and adds zone support.
• Aligns to the existing host-managed models defined in the ZAC/ZBC specifications.
  - Note that it does not map 1:1. Beware of the details.
• Optimized for Solid State Drives
  - Zone Capacity
  - Zone Attributes introduced to optimize for SSD characteristics
  - Zone Append
  - Zone Descriptors
Host-Managed Zoned Block Devices

• Zone States
  – Empty, Implicitly Opened, Explicitly Opened, Closed, Full, Read Only, and Offline.
  – Changes state upon writes, zone management commands, and device resets.

• Zone Management
  – Open Zone, Close Zone, Finish Zone, and Reset Zone

• Zone Size & Zone Capacity
  – Zone Size is fixed
  – Zone Capacity is the writeable area within a zone
Zone Excursions & Variable Zone Size

• For NVMe devices that implement the Zoned Command Set, there is optional support for:
  – Variable Zone Capacity
    • The completion of Reset Zone command may result in a notification that zone capacity has changed.
  – Zone Excursions
    • The device can transition a zone to Full before writes reach the Zone Capacity. Host will receive an AEN and write failure if writing after the transition.

• If device implements, the host shall implement as well
  – Incoherent state model if not – Software must be specifically written to understand that zone capacity can change.

-- Zone Excursion

Zone Size (e.g., 512MB)          Zone Start LBA
------------------------------------------
Zone X - 1                                      Zone X
                                                  Zone Capacity (E.g., 500MB)
                                                  Zone X + 1
Zoned Namespaces TP

Zone Append

• ZAC/ZBC requires strict write ordering
  – Limits write performance, increases host overhead

• Low scalability with multiple writers to a zone
  – One writer per zone -> Good performance
  – Multiple writers per zone -> Lock contention

• Can improve by writing multiple Zones, but performance is limited

• With Zone Append, we scale
  – Append data to a zone with implicit write pointer
  – Drive returns LBA where data was written in zone

![Graph showing performance comparison between 1 Zone, 4 Zones, and Zone Append]
Zone Descriptors

• Fixed-sized data associated to zone
  – Upon opening a zone - associate a fixed-sized amount of data.
  – Invalidated upon zone reset.
  – Same size for all zones.

• Use-case
  – Out-of-band recovery path by allocating UUID’s to each zone.
  – Zones can be self-identifying.
ZNS: Synergies w/ ZAC/ZBC software ecosystem

- Exposed as Zoned Block Devices (ZBD)
- Reuse existing work already done for ZAC/ZBC devices
- Existing ZBD-aware file systems & device mappers “just work”
  - Few additions to support to ZNS
- Integrates with file-systems and applications
  - RocksDB, Ceph, fio, libzns, ...
- ZAC/ZBC devices are already in production at technology adopters and a mature storage stack is available through the Linux® ecosystem.

* = Enhanced data paths for SMR/ZNS drives
ZNS Support in Linux

Shows up as a host-managed Zoned Block Device
Fio Workload
Write sequentially to zones
RocksDB Workload – End-to-end Optimized

• Conventional Enterprise Drive: 90% Full – 5x Write Amplification (WA)
• ZNS Drive: 90% Full
  • **1x WA** – 5x more throughput or increase lifetime 5x lifetime

![ZNS NVMe SSD Development Platform](image-url)
Zoned Namespaces

Coming to a Data Center Near You!

ZonedStorage.io
Documentation, Getting Started, Application Support and much more